SIMPLE AND MULTIPLE LINEAR REGRESSION (OLS, Dummy Variable trap)

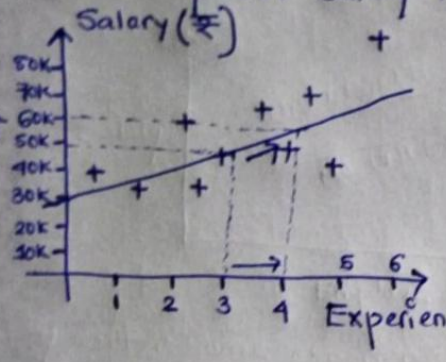
SIMPLE LINEAR REGRESSION,

$$y = b_0 + b_1 x_1$$

b_0 → Constant

y → Dependent Variable (DV), x_1 → Independent variable (IV), b_1 → Coefficient of x_1

Example – How a person salary is dependent on Experience.



Salary = $b_0 + b_1$ (Experience)

b_0 → It is constant where the line crosses the vertical axis. = 30k

Inference, when Experience is 0, then the average starting salary is ₹30k. i.e, Fresher salary.

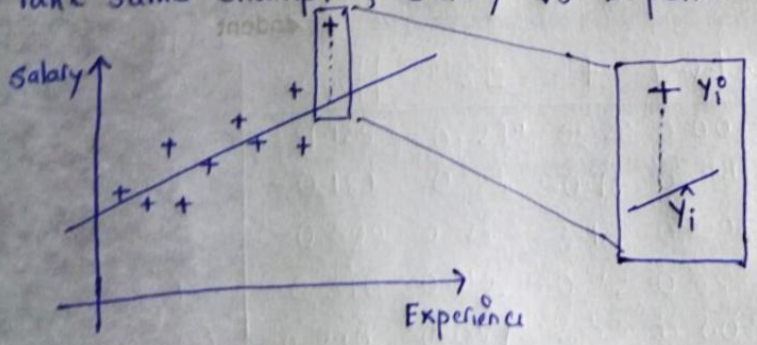
b_1 → the value of slope in unit change of x / Experience. Suppose we want to see how much salary change is there if experience is increased by one year. Inference, if the experience is increased from 3 to 4, then salary increased by ₹ 10k i.e, 50k to 60k.

So if we want to build the eqn, Salary = 30,000 + b_1 (Experience)

$b_1 \neq 10k$ because that is change in y, but b_1 is the value of slope.

One of the method to find best fit line, is **Ordinary Least Square (OLS)**

Take same example, Salary Vs Experience.



y_i → Actual Salary.

$\hat{y}_i$ → Predicted Salary

$(y_i - \hat{y}_i)^2$ → Error.

Then for this regression line, we find the total error. → Sum $(y - \hat{y})^2$.

→ So for this regression line, we will calculate the total error. $\text{Sum}_1 (y - \hat{y})^2$.

→ We will draw another line, and for that line we will calculate some $\text{Sum}_2 (y - \hat{y})^2$.

→ We will choose the minimum $\text{Sum}_n (y - \hat{y})^2$, which will be the best line to find fit.

This method is known as Ordinary Least Square.

→ Multiple Linear Regression is on next page.

Simple linear regression, $y = b_0 + b_1 x_1$

Multiple linear regression, Dependent variable (DV) $y = b_0 + b_1 * x_1 + b_2 * x_2 + \dots + b_n x_n$ (Independent variables (IV))

b_0 ← Constant; $b_1, b_2, \dots, b_n$ ← Coefficients

Assumption of linear regression –
i) Linearity
ii) Homoscedasticity
iii) Multivariate normality
iv) Independence of errors
v) Lack of multicollinearity

Dummy Variables → Suppose we have a categorical column which consist of set of alphabet values then we need to change the alphabets to numeric. That conversion is known as dummy variable.

Dummy Variable Trap

Profit (Dependent Variable)	Marketing (Independent Variable)	State/City
2000	1000	Pune
3000	1500	Mumbai
4000	1300	Mumbai
2500	1700	Pune

Dummy Variables →

Profit	Marketing	Pune	Mumbai
2000	1000	1	0
3000	1500	0	1
4000	1300	0	1
2500	1700	1	0

$(D_2 = 1 - D_1)$

Dummy Variable Trap – Pune & Mumbai columns are mirror image/switches. If one column is 1, other will be 0. and vice versa. So, we should remove one column.
This issue can be raised under multicollinearity. And in assumption, we said there should be no multicollinearity.

5 methods of building regression models –
1. All in variable
2. Backward elimination
3. Forward selection
4. Bidirectional Elimination
5. Score comparision

(2, 3, 4 → Stepwise regression)

Some only bidirectional elimination is only refer as stepwise regression.

1. **All in** → Choose all the variable first and start removing wasteful variable.
2. **Backward elimination** → ① Select a significance level to stay in model (SL = 0.05)
② Fit the model with all predictor variables.
③ Consider the predictor with highest Pvalue. IF P > SL, go ④
④ Remove the predictor.
⑤ Fit the model without this variables. (Final model)
3. **Forward Selection** – ① Select a significance level to enter the model (SL = 0.05)
② Fit all simple regression model $y \sim x_n$, select one with lowest P value.
③ Keep this variable and fit all possible models with one extra predictor added to one(s) already have.
④ Consider the predictor with lowest P-value. (Final model)
4. **Bidirectional elimination** – ① Select a significance level to enter and stay in a model.
② Perform the next step of forward selection ③ Perform all backward elimination steps
④ No new variable can enter (select) or delete (backward) Final model
5. **Score Comparision** – Adjusted R square, Pvalue, Aic, BIC.